Nakala. Un entrepôt de données pour les SHS

Victoria Le Fourner, Florence Perret

Nakala

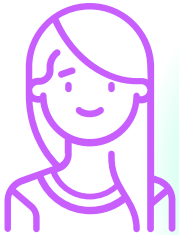
Un entrepôt de données pour les SHS

Florence PERRET (CNRS)
Victoria LE FOURNER (CNRS)

Université de Lille

MESHS Lille Nord de France

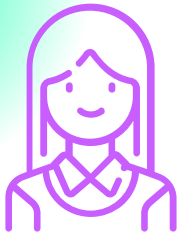
Les intervenantes



Florence Perret

Ingénieure d’études

florence.perret@univ-lille.fr

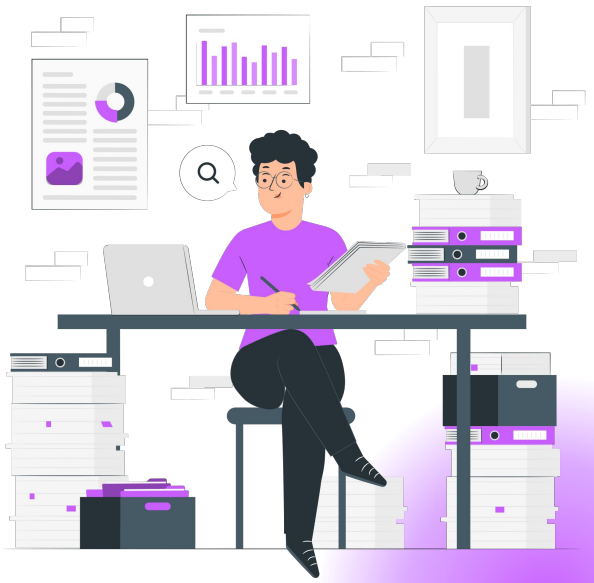


Victoria Le Fourner

Ingénieure d’études

victoria.lefourner@univ-lille.fr

Table of contents



01

Science ouverte et SHS

Contexte théorique et institutionnel

Faire de la recherche en SHS

Des données hétérogènes

Des projets de recherche sur un temps long

Des besoins précis, adaptés aux types de données

Que sont les données de la recherche pour vous ?

a) Des tableurs dans lesquels sont organisées les résultats d'enquêtes, d'observations, de relevés
b) Des documents décrivant la structure des données (modèles)
c) Une bibliographie
d) Un dossier regroupant des images, des vidéos, des sons ou des textes

Les données de la recherche
définition

Enregistrements factuels (chiffres, images, sons, vidéos…), qui sont utilisés comme sources principales pour la recherche scientifique et sont également reconnus par la communauté scientifique comme nécessaires pour valider les résultats de recherche.

OCDE, *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*, 2007

Le cycle de vie des données de la recherche

Le cycle de vie des données

1. Conception du projet
2. Création des données
3. Traitement des données
4. Analyse des données
5. Archivage et conservation
6. Diffusion des données et métadonnées
7. Réutilisation des données

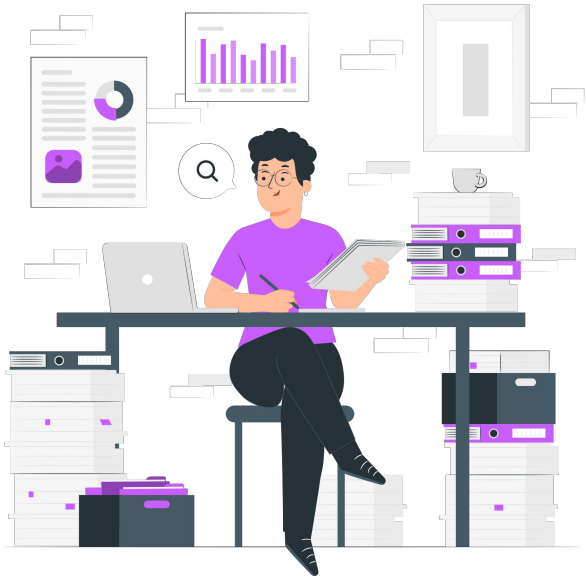
Conception du projet

Réflexion sur les données

Réflexion sur les DCP

Rédaction du PGD

Création des données



Collecte et acquisition de données

Création des premières métadonnées

Déclaration du traitement des DCP

Traitement des données

Préparation des données brutes

Nettoyage, vérification, curation

Stockage et échanges sécurisés

Analyse des données

Utilisation d’outils et méthodes de production de résultats

Mise en conformité des DCP si modification des analyses

Archivage et conservation

Réflexion sur la pertinence de la conservation

Organiser les jeux de données et métadonnées associées

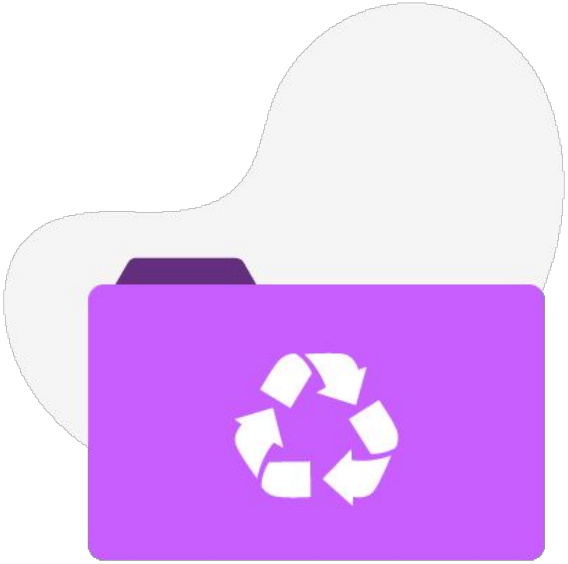
Diffusion des données et métadonnées

Dépôt dans un entrepôt, catalogue de données

Attribution de DOI

Attribution de licences

Réutilisation des données



S’informer sur les données et leurs licences

Évaluer la réutilisation de ses données

Huma-Num, une infrastructure pour les SHS

Suis-je éligible à l’offre de services?

Monde académique et projet scientifique validé par Huma-Num.

Pour avoir un projet scientifique validé, plusieurs critères possibles :

- besoins spécifiques sur le traitement des données
- engagement sur l’interopérabilité des données de la recherche et des métadonnées associées
- démarche d’archivage à long terme des données.

L’Infrastructure de Recherche* Huma-Num

Infrastructure de recherche financée par le Ministère de l’ESR dédiée à la **gestion des données en SHS**.

- Des services pour les données
- Des Consortiums
- Le HN Lab
- Coordination des communautés européennes et internationales
- Maillage thématique et géographique du territoire

Huma-Num
la TGIR des humanités numériques


DISSÉMINATION
PUBLICATION
SIGNALEMENT
ANIMATION
COMMUNAUTÉS
NATIONALES &
INTERNATIONALES
HN
HN Lab
OUTILS COLLABORATIFS
STOCKAGE
TRAITEMENT
QUALITÉ DES DONNÉES
ACCOMPAGNEMENT

Les missions d’Huma-Num

Vision

Développer l’appropriation du cycle de vie des données numériques

Organisation

Organiser et faciliter le travail en équipe et la gestion de projets

Développer

Proposer des services pour les données au juste niveau et au bon moment

Compréhension

Travailler collectivement autour des données

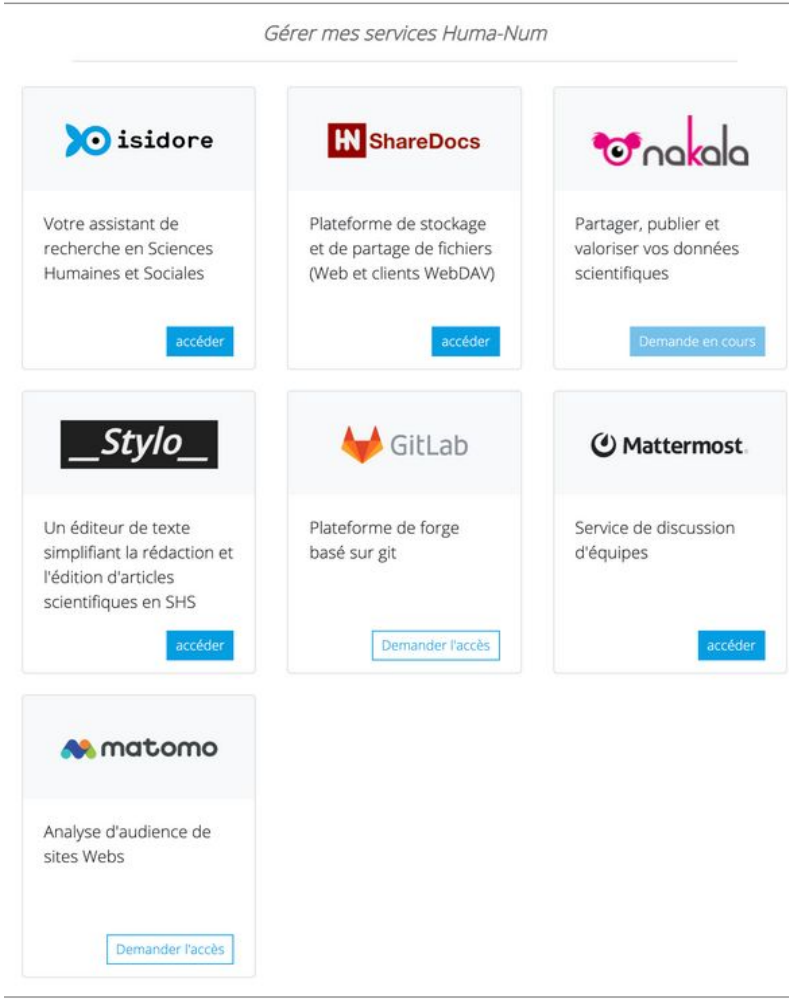
Accès

Stabiliser l’accès aux données et aux outils

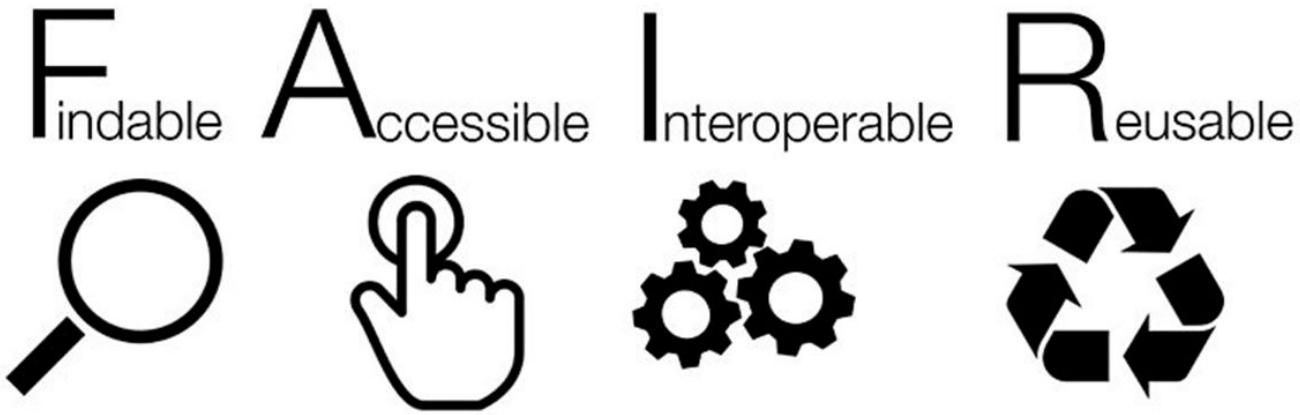
Partenariat

Co créer des services

Accéder aux services



Les principes FAIR



ORGANISATION

Des services pour organiser le travail collaboratif autour de vos données.

- ShareDocs
- GitLab
- Kanboard
- Mattermost

TRAITEMENT

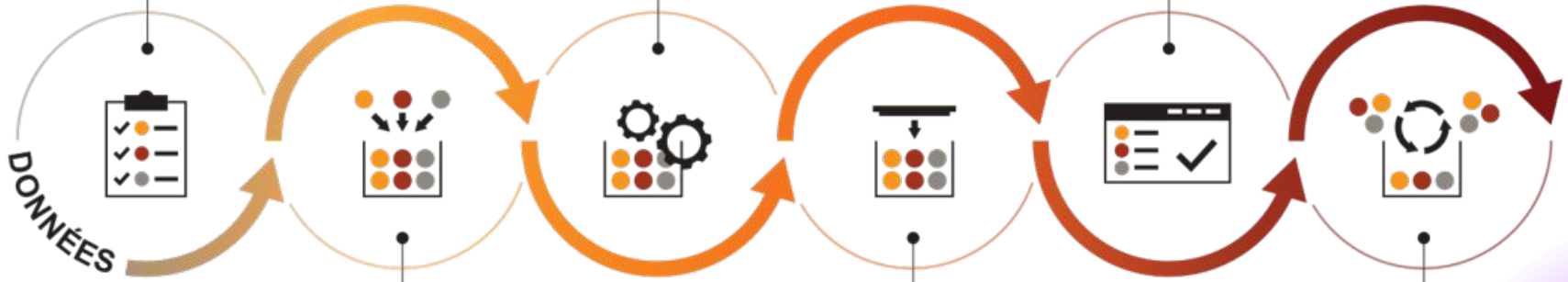
Des services et outils spécifiques pour le traitement et l'analyse de vos données.

- Calcul statistique et environnements R
- Logiciels d'enquête et d'analyse de données
- Reconnaissance de caractères
- Puissance de calcul (+ CC-IN2P3)
- ...

PUBLICATION

Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

- Hébergement Web
- Machines Virtuelles
- Nakala
- Isidore



COLLECTE

Des services de stockage sécurisé pour la collecte et la création de vos données.

- ShareDocs
- Huma-Num Box

PRÉSERVATION

Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

- Nakala
- Huma-Num Box
- Préservation à long terme (+ CINES)

RÉUTILISATION

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

- Portail web
- API
- Triplestore
- OAI-PMH

Nakala ?

Un entrepôt

Dépôt, description, conservation, recherche et diffusion de jeux de données

Sécurisé

Serveurs en France (IN2P3)

Technologies éprouvées

Dédié aux SHS

Proche de la communauté SHS

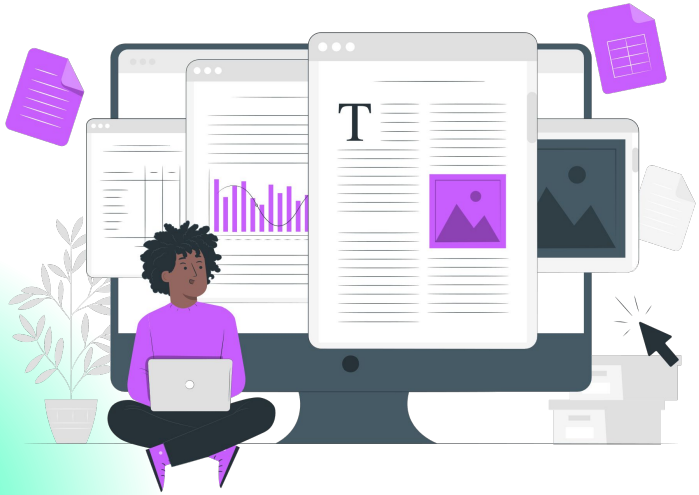
Groupes de Travail pour répondre aux attentes

FAIR

Création de DOI (reusable), visionneuse (accesible), respect des standards (interoperable), moissonné par Europeana, Gallica et Isidore (findable)

02

Les données sur Nakala



Interface, standard de métadonnées et vocabulaire

Un peu de théorie

Métadonnées et licence

Les métadonnées
Pourquoi ?

Rendre intelligible

Pour vous ou pour les autres !

Rendre recherchable

Sur un moteur de recherche

Permettre un aperçu

Savoir rapidement si ces données vous intéressent

Répondre aux exigences de la Science Ouverte

Les métadonnées
Qu’est-ce que c’est ?

Structure simple :

label = valeur du label

Titre = La Curée

Structure complexe :

langueDuTitre = fr

typeDeDonnéesDuTitre = chaîne de caractères

Le Dublin Core

Format de métadonnées créé en 1995 (Dublin, USA)

Cœur : Dublin Core simple, 15 éléments facultatifs et répétables

Dublin Core qualifié / enrichi : environ 55 termes dont les 15 éléments du DB simple + audience, provenance and rightsHolder + termes qui viennent préciser les propriétés simples de manière plus fine

Exemple de la relation DC simple / qualifié

DC SIMPLE	DC QUALIFIÉ	Définition
dc:date	dcterms:created	Date de création de la ressources
	dcterms:issued	Date de sortie officielle
	dcterms:modified	Date à laquelle la ressource a été modifiée

Les règles générales du DC

Pour commencer, consulter la documentation :
https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

1.

Tous les termes sont répétables (ex : plusieurs auteurs)

2.

L'usage de vocabulaires contrôlés est encouragé

3.

Il est préférable d'utiliser le DC ÉTENDU plutôt que le DC SIMPLE

4.

Il est possible de créer son propre système normatif à partir du DC

Pour en savoir plus : Laurent Capelli et Aurelia Vasile, "Exposer les données d'un projet de recherche avec NAKALA et NAKALA_Press", Tuto@Mate, 3/12/2022
https://mate-shs.cnrs.fr/actions/tutomate/tuto39-nakala-capelli-vasile/

Le Dublin Core et Nakala

Les cinq éléments obligatoires

	nakala:title	nakala:type	nakala:creator	nakala:created	nakala:licence
Doc 1	Chaîne de caractères	Image	Anonyme	1933	LO
Doc 2	Chaîne de caractères	Fonds d'archives	Nom Prénom	Inconnue	CC-BY-SA
Doc 3	Chaîne de caractères	Dataset	Nom Prénom	1934-06-25	Etalab Open License 2.0

Les licences

Une licence de diffusion est un instrument juridique, complémentaire au droit d'auteur. Elle permet au titulaire des droits sur une œuvre d'accorder à l'avance aux utilisateurs certains droits d'utilisation de cette œuvre.

Pour favoriser la réutilisation des jeux de données que vous rendez publics, privilégiez des licences largement utilisées.

Quelques exemples
Les licences Creative Commons (CC-BY | CC-BY-SA | CC0)
La licence ouverte (LO)
Les licences de l'Open Knowledge Foundation (OKF)
ODC-by (Open Database Commons)
ODC-ODBL (Open database License)
PDDL (Public domain dedication and license)

Dedieu L. ; Fily M.F. 2015. Rendre publics ses jeux de données scientifiques, en 6 points. Montpellier (FRA) : CIRAD, 6 p. https://doi.org/10.18167/coopist/0059

Entrer dans Nakala

Donnée, Collections, Rôles, Accès, Interface

Nakala est un service d'Huma-Num permettant à des chercheurs, enseignants-chercheurs et équipes de recherche de partager, publier et valoriser tous types de données numériques documentées (fichiers textes, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé afin de les publier en accord avec les principes du *FAIR data* (Facile à trouver, Accessible, Interopérable et Réutilisable).

Documentation Huma-Num

Une donnée dans Nakala

Une donnée sur Nakala : 1 à N fichiers

Deux statuts :

- Déposée : visibilité restreinte, état transitoire, espace de stockage limité
- Publiée : accessible, réutiilsable, attribution d'un DOI, pas de limite de stockage, pas de possibilité de supprimer

Les collections

Une collection regroupe un ensemble de données cohérentes.

Une donnée peut appartenir à plusieurs collections.

Il n’y a pas de hiérarchie entre les collections.

Les rôles

Déposant

ROLE_DEPOSITOR

Propriétaire

ROLE_OWNER

Administrateur

ROLE_ADMIN

Éditeur

ROLE_EDITOR

Lecteur

ROLE_READER

Anonyme

GUEST

Contrôler l’accès

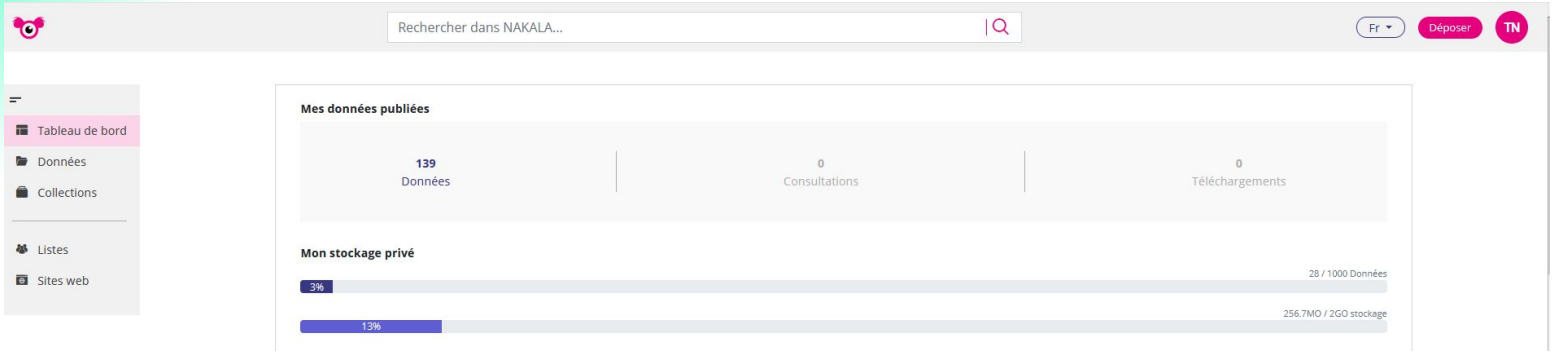
Les droits associées aux fichiers d’une donnée dépendent des critères suivants :

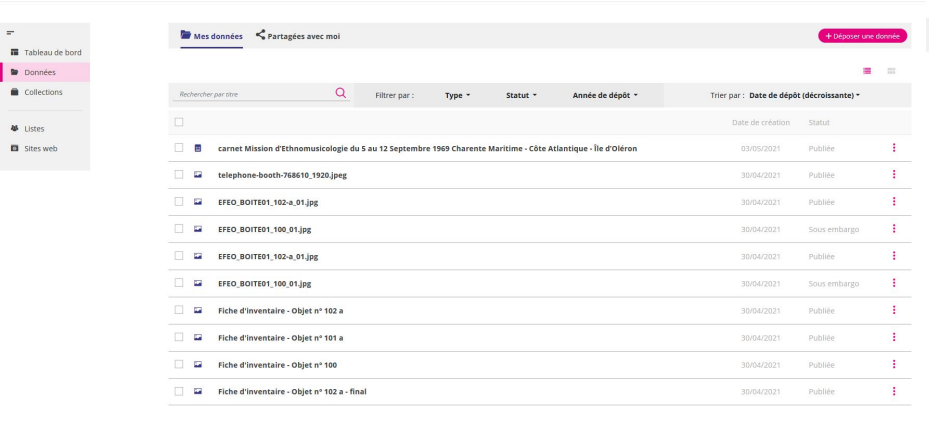
- le rôle de l’utilisateur sur la donnée (cf. plus haut)
- le statut de la donnée (déposée, publiée, ancienne version, supprimée)
- la présence ou non d’une date d’embargo en cours sur le fichier

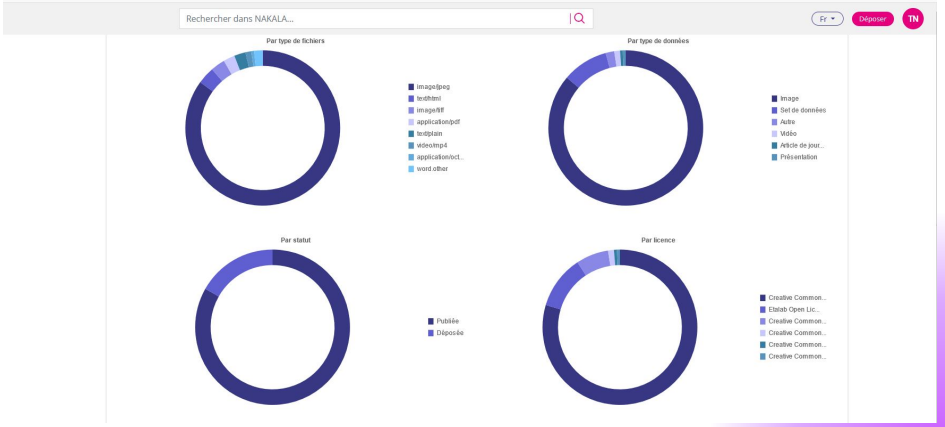
Un embargo
définition

Période pendant laquelle les articles et les données de recherche déposés dans un réservoir ne sont pas accessibles librement.

Vue du Tableau de bord et des données

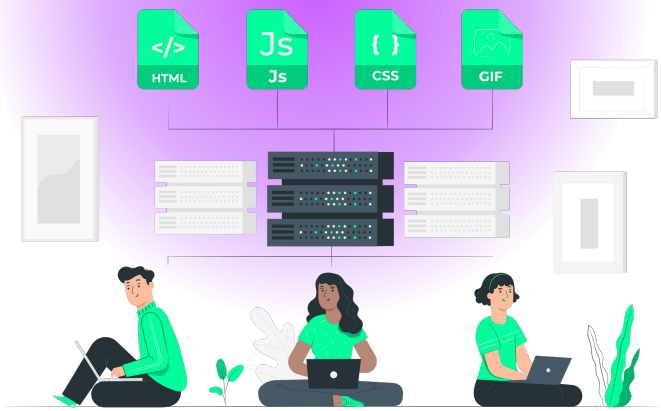






Organisation de l’interface

- Un onglet **Tableau de bord** permet de suivre les métriques de vos données de façon chiffrée : nombre de données déposées/publiées, consultations de ces données, téléchargements, stockage privé, fréquence des dépôts. Cette page vous est personnelle ;
- Un onglet **Données** permet de lister vos données ou des données que d’autres utilisateurs ont partagées avec vous ;
- Un onglet **Collections** permet de gérer vos différentes collections ainsi que celles partagées avec vous par d’autres utilisateurs ;
- Un onglet **Listes** permet la gestion de vos groupes d’utilisateurs afin de gérer plus facilement les droits d’accès et de contribution à vos dépôts ;
- Un onglet **Sites web** permet la gestion de vos différents sites Nakala_Press.



03

Pour aller plus loin avec Nakala

Éditorialiser ses données

Nakala_Press

Interface, Enjeux

Éditorialiser avec Nakala_Press

1. Collection publique

2. Données publiées

Une collection = Un site

https://votre_site.nakala.fr

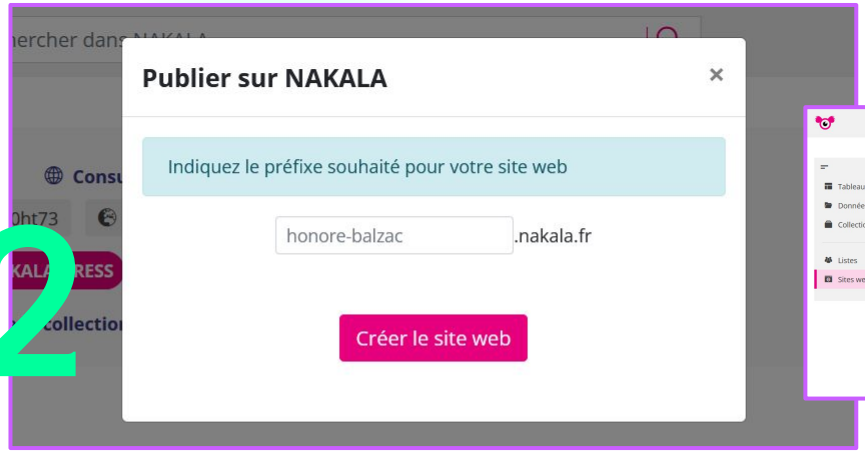
01


Rechercher dans NAKALA...
Fr
Honore_Balzac
Consulter
ID : 10.34847/nkl.c3a0ht73
Publique
Supprimer
Créer un site NAKALA_PRESS
Afficher les détails de la collection
Rechercher par titre
Filtrer par :
Type
Statut
Année de dépôt
Trier par :
Date de dépôt
Statut
test
10/05/2022
Publiée

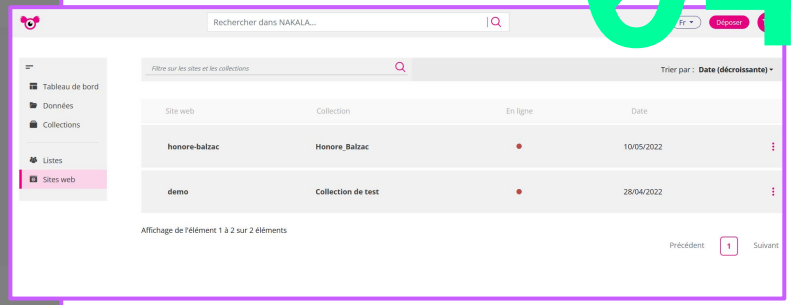
03


Honore_Balzac
Consulter
ID : 10.34847/nkl.c3a0ht73
Publique
Le site web est publié et consultable sur https://honore-balzac.nakala.fr
Afficher les détails de la collection
Consulter
Administrer
Supprimer

02

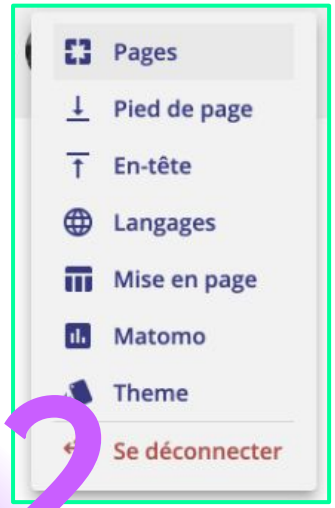

Publier sur NAKALA
Indiquez le préfixe souhaité pour votre site web
honore-balzac
.nakala.fr
Créer le site web

04

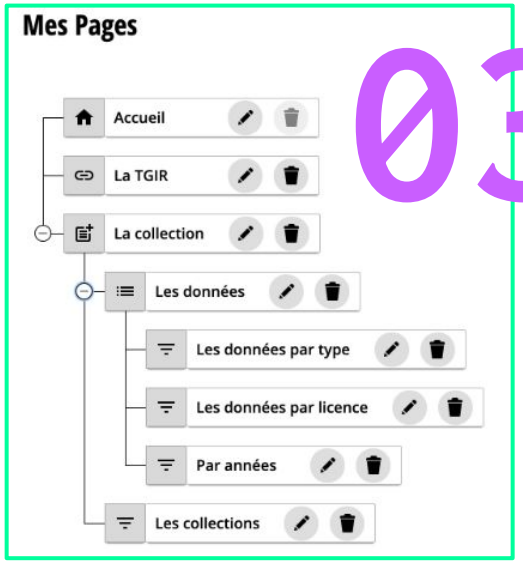

Rechercher dans NAKALA...
Déposer
Tableau de bord
Données
Collections
Listes
Sites web
Trier par : Date (décroissante)
Site web
Collection
En ligne
Date
honore-balzac
Honore_Balzac
10/05/2022
demo
Collection de test
28/04/2022
Affichage de l'élément 1 à 2 sur 2 éléments
Précédent
Suivant

Configuration pas à pas

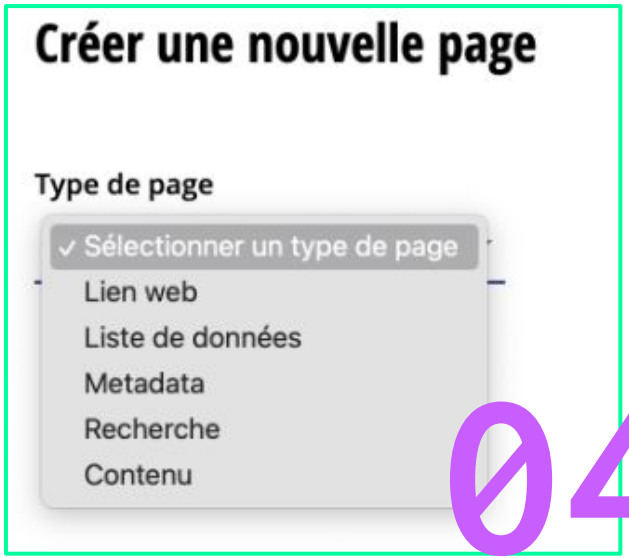
01



02



03



04

Rechercher et Exposer

Isidore, OAI PMH, Triple Store et autre API...

Isidore

Moteur de recherche permettant de découvrir et de trouver des publications, des données numériques et profils de chercheurs et chercheures en SHS.

Rechercher des données

API Nakala

Une API est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

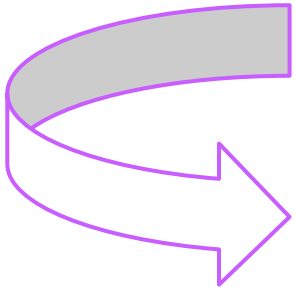
OAI PMH

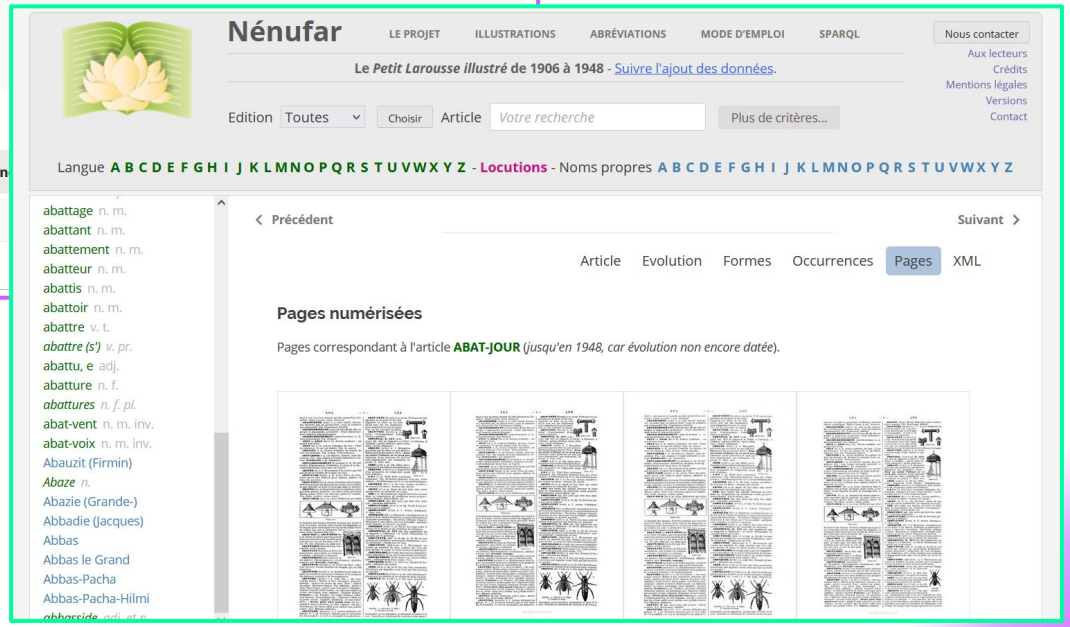
Protocole permettant d’ échanger sur Internet des métadonnées entre plusieurs institutions, afin de multiplier les accès aux documents numériques.

Triple Store

Base de données qui ne contient que des triplets RDF.
Actuellement en cours de mise à jour.

Exposer différemment





En résumé, Nakala c’est quoi ?

Que retenir?

NAKALA

Pour les SHS

Répond aux besoins spécifiques des SHS

Sécurisé

Répond aux exigences de sécurité de l'ouverture des données

FAIR

Répond aux principes FAIR

Science Ouverte

Répond aux exigences du Plan National pour la Science Ouverte

Huma-Num

Maintien Nakala depuis 2014

Exposable

Permet de créer rapidement un site pour exposer ses données

Que fait Nakala ?

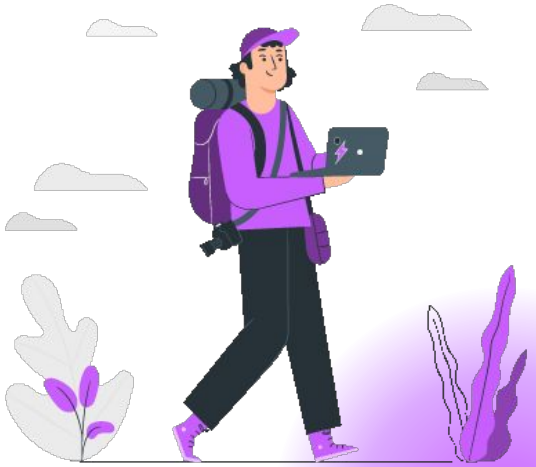
- Il vous décharge de la gestion de vos données ;
- Il vous permet de les visualiser ;
- Il vous permet de les regrouper et les présenter dans des collections homogènes ;
- Il prend en charge le partage interopérable des données et des métadonnées et leur citabilité ;
- Il dissocie le stockage de données de leur présentation ;
- Il prépare le référencement des données dans ISIDORE et facilite le processus d'archivage à long terme ;
- Il permet l'éditorialisation de vos données dans un site web personnalisé de type *https://monprojet.nakala.fr* grâce au module de publication Nakala_Press.

Que ne fait pas Nakala ?

- Il n’enrichit pas les données ;
- Il ne permet pas un stockage des données à caractères sensibles ou sous-droits.

04

Prise en main de Nakala



Expérimentation dans le bac à sable

Bac à sable

https://test.nakala.fr/

Connectez-vous avec les identifiants tests

HummanID	Mot de passe	Clé d'API
tnakala	IamTesting2020	01234567-89ab-cdef-0123-456789abcdef
unakala1	IamTesting2020	33170cfe-f53c-550b-5fb6-4814ce981293
unakala2	IamTesting2020	f41f5957-d396-3bb9-ce35-a4692773f636
unakala3	IamTesting2020	aae99aba-476e-4ff2-2886-0aaf1bfa6fd2

Mise en situation

Vous avez été recruté pour le projet ANR PERLE sur la circulation des estampes de mode.

Le projet s’intéresse aux pays qui ont adopté la mode française de la robe de mariée blanche et étudie la presse de mode.

Créer des collections

- Collection par pays
- Collection par journal de modes
- Collection pour le projet

Déposer une donnée et un fichier

Une ligne par personne en fonction de votre numéro.

Que remarquez-vous ?

Il y a t-il des difficultés en fonction des métadonnées préparées?

Déposer en multi-fichiers

Les chercheurs du projet ne vous ont pas facilité la tâche et les fichiers qu'ils ont utilisés ne sont pas prêts pour le dépôt.

Dans chaque dossier, comment pouvez-vous assurer le dépôt en enrichissant les données pour être sûr qu'elles soient réutilisées?

Quelles métadonnées remplir ?

1	**Type**
2	**Titre**
3	**Auteur**
4	**Date de création**
5	**Licence**
6	Description
7	Mots-Clés
8	Langues
9	Contributor
10	Provenance

Pour en savoir plus :

https://documentation.huma-num.fr/nakala/

Merci !

Avez-vous des questions ?

Nous restons à votre écoute à la MESHS !